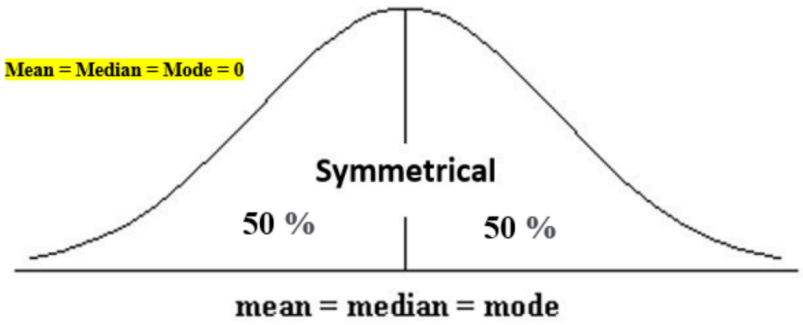
Skewness

If the values of a specific independent variable (feature) are skewed, depending on the model, skewness may violate model assumptions or may reduce the interpretation of feature importance.

In statistics, skewness is a degree of asymmetry observed in a probability distribution that deviates from the symmetrical normal distribution (bell
curve) in a given set of data.

The normal distribution helps to know a skewness. When we talk about normal distribution, data symmetrically distributed. The symmetrical distribution has zero skewness as all measures of a central tendency lies in the middle.

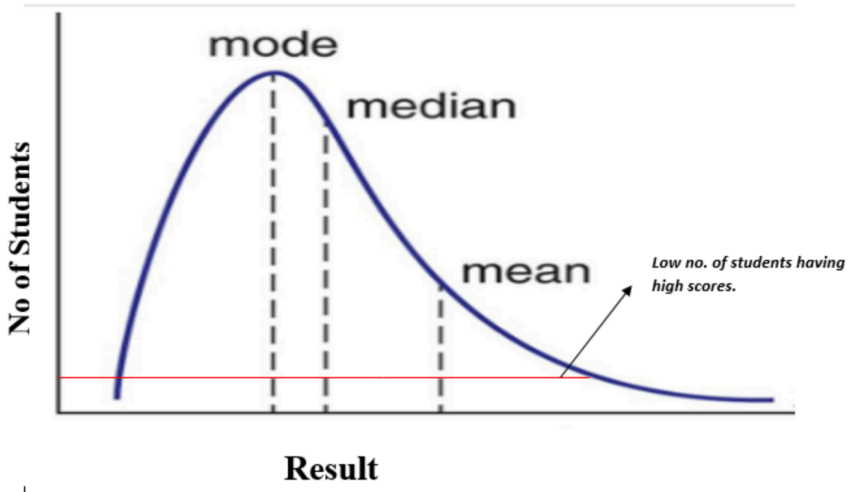


When data is symmetrically distributed, the left-hand side, and right-hand side, contain the same number of observations. (If the dataset has 90 values, then the left-hand side has 45 observations, and the right-hand side has 45 observations.). But, what if not symmetrical distributed? That data is called asymmetrical data, and that time skewness
comes into the picture.

Types of skewness

1. Positive skewed or right-skewed

In statistics, a positively skewed distribution is a sort of distribution where, u*nlike symmetrically distributed data where all measures of the central tendency (mean, median, and mode) equal each other,* with positively skewed data, the measures are dispersing, which means Positively Skewed Distribution is a type of distribution where the mean, median, and mode of the distribution are positive rather than negative or zero.



Mean > Median > Mode

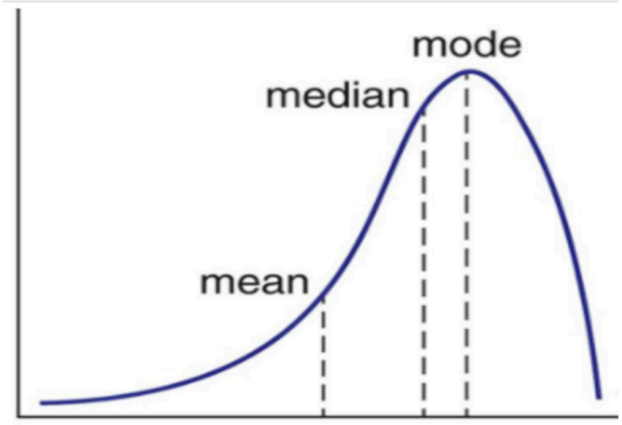
In positively skewed, the mean of the data is greater than the median (a large number of data-pushed on the right-hand side). In other words, the results are bent towards the lower side. The mean will be more than the median as the median is the middle value and mode is always the highest value

The extreme positive skewness is not desirable for distribution, as a high level of skewness can cause misleading results. The data transformation tools are helping to make the skewed data closer to a normal distribution. For positively skewed distributions, the famous transformation is the log transformation. The log transformation proposes the calculations of the natural logarithm for each value in the dataset.

2. Negative skewed or left-skewed

A negatively skewed distribution is the straight reverse of a positively skewed distribution. In statistics, negatively skewed distribution refers to the distribution model where more values are plots on the right side of the graph, and the tail of the distribution is spreading on the left side.

In negatively skewed, the mean of the data is less than the median (a large number of data-pushed on the left-hand side). Negatively Skewed Distribution is a type of distribution where the mean, median, and mode of the distribution are negative rather than positive or zero.



Mode > Median > mean

Calculate the skewness coefficient of the sample

Pearson's first coefficient of skewness

Subtract a mode from a mean, then divides the difference by standard deviation.

$$\text{Pearson's first coefficient} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

As Pearson's correlation coefficient differs from -1 (perfect negative linear relationship) to +1 (perfect positive linear relationship), including a value of 0 indicating no linear relationship, When we divide the covariance values by the standard deviation, it truly scales the value down to a limited range of **-1 to +1.** That accurately the range of the correlation values.

Pearson's first coefficient of skewness is helping if the data present high mode. But, if the data have low mode or various modes, Pearson's first coefficient is not preferred, and Pearson's second coefficient may be superior, as it does not rely on the mode.

Pearson's second coefficient of skewness

Multiply the difference by 3, and divide the product by standard deviation.

$$\text{Pearson's second coefficient} = \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

$$\text{Mean} - \text{Mode} \approx 3\,(\text{Mean} - \text{Median})$$

If the skewness is between -0.5 & 0.5, the data are nearly symmetrical.

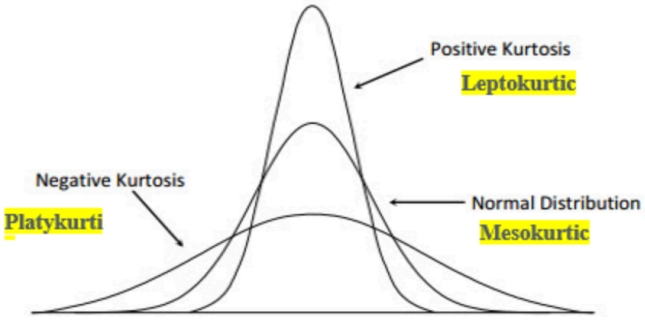
If the skewness is between -1 & -0.5 (negative skewed) or between 0.5 & 1(positive skewed), the data are slightly skewed.

If the skewness is lower than -1 (negative skewed) or greater than 1 (positive skewed), the data are extremely skewed.

Kurtosis

Kurtosis refers to the degree of presence of outliers in the distribution.

Kurtosis is a statistical measure, whether the data is heavy-tailed or light-tailed in a normal distribution.



In finance, kurtosis is used as a measure of financial risk. A large kurtosis is associated with a high level of risk for an investment because it indicates that there are high probabilities of extremely large and extremely small returns. On the other hand, a small kurtosis signals a moderate level of risk because the probabilities of extreme returns are relatively low.

Excess Kurtosis

The excess kurtosis is used in statistics and probability theory to compare the kurtosis coefficient with that normal distribution. Excess kurtosis can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near to zero (Mesokurtic distribution). Since normal distributions have a kurtosis of 3, excess kurtosis is calculating by subtracting kurtosis by 3.

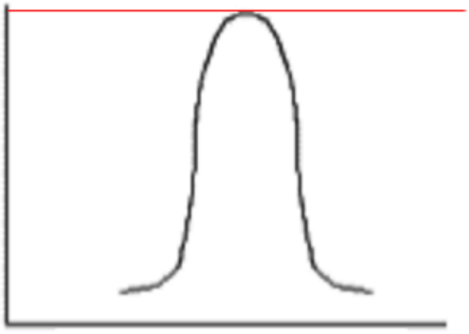
Excess kurtosis = Kurt – 3

Types of excess kurtosis

1. *Leptokurtic or heavy-tailed distribution (kurtosis more than normal distribution).*
2. *Mesokurtic (kurtosis same as the normal distribution).*
3. *Platykurtic or short-tailed distribution (kurtosis less than normal distribution).*

Leptokurtic (kurtosis > 3)

Leptokurtic is having very long and skinny tails, which means there are more chances of outliers. Positive values of kurtosis indicate that distribution is peaked and possesses thick tails. An extreme positive kurtosis indicates a distribution where more of the numbers are located in the tails of the distribution instead of around the mean.



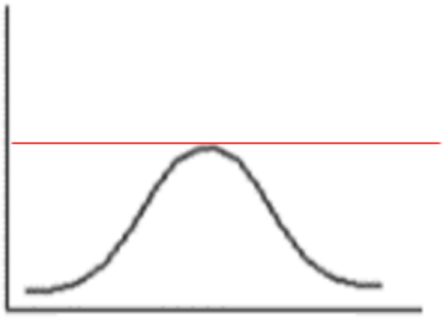
Leptokurtic

platykurtic (kurtosis < 3)

Platykurtic having a lower tail and stretched around center tails means most of the data points are present in high proximity with mean. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution.

Mesokurtic (kurtosis = 3)

Mesokurtic is the same as the normal distribution, which means kurtosis is near to 0. In Mesokurtic, distributions are moderate in breadth, and curves are a medium peaked height.



Mesokurtic

Mesokurtic = 3 – 3 = 0

Summary

The skewness is a measure of symmetry or asymmetry of data distribution, and kurtosis measures whether data is heavy-tailed or light-tailed in a normal distribution. Data can be positive-skewed (data-pushed towards the right side) or negative-skewed (data-pushed towards the left side).

When data skewed, the tail region may behave as an outlier for the statistical model, and outliers unsympathetically affect the model's performance especially regression-based models. Some statistical models are hardy to outliers like Tree-based models, but it will limit the possibility to try other models. So there is a necessity to transform the skewed data to close enough to a Normal distribution.

Excess kurtosis can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near to zero (Mesokurtic distribution). *Leptokurtic distribution (kurtosis more than normal distribution).Mesokurtic distribution (kurtosis same as the normal distribution).Platykurtic distribution (kurtosis less than normal distribution).*